15.5 A 52.4mW 3D Graphics Processor with 141Mvertices/s Vertex Shader and 3 Power Domains of Dynamic Voltage and Frequency Scaling

Byeong-Gyu Nam, Jeabin Lee, Kwanho Kim, Seung Jin Lee, Hoi-Jun Yoo

KAIST, Daejeon, Korea

Real-time 3D graphics have been widely adopted on handheld devices such as cellular phones and PDAs with high performance processors consuming only limited power [1-3]. Recently, a vertex shader of 120Mvertices/s and 106mW at 60fps was presented [3], but it consumed a large silicon area of 1.5M transistors to integrate 16 floating-point multipliers for the fast matrix multiplications required. A full 3D graphics pipeline processor is presented with 141Mvertices/s vertex shader (VS) and 52.4mW power consumption at 60fps. The VS utilizes a logarithmic number system (LNS) for high speed and small area with only 968k transistors. For low power consumption, the 3 power domains of the chip are separately controlled by dynamic voltage and frequency scaling (DVFS).

Figure 15.5.1 shows the overall block diagram of the 3D graphics processor. It consists of an ARM10-compatible 32b RISC processor, a 128b programmable VS, a rendering engine (RE) and 3 power management units (PMUs). The RISC works as the main processor and runs application programs generating the geometry transformation matrices for objects in a scene. The VS transfers the matrix from the matrix FIFO to its constant memory (CMEM) and then performs various geometry operations on the fetched vertices by executing a vertex program in instruction memory (IMEM). The RE fills up the internal pixels of the triangle of which vertices come from the vertex cache (VC). The chip is divided into 3 different power domains; RISC, VS and RE. The 4KB matrix FIFO and 64B index FIFO are used for buffering the matrices and vertex indices, respectively, to keep the pipeline throughput stable despite the response delay of the PMUs. In these FIFOs, gray coding is used for the read/write pointer for reliable operation between different read/write clock domains. Between power domains, level shifters and synchronizers are inserted to adjust signal levels and avoid the metastability of transferred data. The 4KB physical entry memory for the index FIFO acts also as a 16-entry transformation and lighting (TnL) VC to reuse the previously processed vertices with 58% hit rate. This results in a single cycle TnL for the vertices in the VC.

The functional units utilizing LNS in [4] were only fixed-point units. However, the proposed VS of Fig. 15.5.2 includes a floating point functional unit that unifies the vector operations (VEC) like vector multiply-and-add, divide, divide-by-square root, dot product and transcendental functions (TRS) like power, logarithm, sine, cosine, arctangent, etc., based on the schemes introduced in [4]. Moreover, it unifies the matrix multiplication of the 4×4 transformation matrix and 1×4 vertex matrix (MAT) with the VEC and TRS in a single 4-way arithmetic platform. Its pipeline achieves a maximum 5-cycle latency and 1-cycle throughput for all these operations except for the MAT, which has a 6-cycle latency and 2-cycle throughput.

Since the 32b×8b Booth multiplier (BMUL) in [4] has the same CSA tree and CPA as the logarithmic converter (LOGC) and antilogarithmic converters (ALOGCs), it can be modified into a programmable BMUL (PMUL) to operate as LOGC for VEC and ALOGC for MAT by just adding 120B look-up tables. As a result, 4 PMULs act as 4 LOGCs or 4 ALOGSs in the E2 stage of Fig. 15.5.2. The geometry transformation in 3D graphics can be calculated by MAT. Since the elements of the transformation matrix for the given objects in a scene are fixed in 3D graphics, they can be pre-converted into LNS before processing of the object. The LOGCs of the E1 stage of Fig. 15.5.2 convert only the vertex matrix to perform MAT. Therefore, a complete MAT requires 16 additions in LNS, 16 ALOGCs and 12 additions of the products in sequence to get the final results. 2 phases are required for the MAT in this VS as shown in Fig. 15.5.2. If the 4 PMULs in the E2 stage are programmed into 4 ALOGCs, 8 ALOGCS are obtained together with 4 ALOGCs in the E3 stage. The 4 CPAs in the E1 stage and 4 CPAs in the E3 stage make a total 8 CPAs for the addition in LNS. The 4 multiplication results from the ALOGCs in the E2 stage and the other 4 from the E3 stage are added in the E4 stage to get the first phase result. With the same process repeated and the accumulation of the first phase result, the final result of the MAT is obtained in the second phase. This results in a 2-cycle throughput for the MAT, which took 4-cycles in [2]. This 2-cycle MAT with the VC achieves a performance of 141Mvertices/s at 200MHz. The average error for this operation is 0.014%, unnoticeable to the naked eye on the small screen used in handheld devices.

In the processor, 3 power domains with separate frequencies and voltages are continuously tuned by tracking workloads. Although an H.264 codec of 2 power domains was reported, it just supported switching between 2 supply voltages and frequencies only in one domain [5]. The workloads of RISC, VS and RE, which operate on objects, vertices and pixels, respectively, can be completely different. Thus, the RISC, VS and RE are in different power domains and their frequencies and voltages are separately controlled according to their workloads. The workloads of RISC and VS are obtained by measuring the occupation levels of the 16-entry matrix FIFO and 32-entry index FIFO, respectively. Each occupation level is compared with the reference level to give *Err* and the value of *Err* is shifted by feedback gain, k_f. A new target frequency, $Fsel_{new}$, is generated by the frequency selection circuit (FS) by subtracting the shifted *Err* from $Fsel_{cur}$ as shown in Fig. 15.5.3. For the RE domain, the software routine running on the RISC measures the time elapsed for the drawing of a scene and sets a new target frequency and voltage adaptively for the next scene. The droop of the FIFO entry level, i.e. high workload due to small polygon, increases the frequency and voltage of the VS to speed up its operation until the FIFO entry level returns to the reference point as shown in Fig. 15.5.4.

The PMU merges a linear regulator in the PLL loop to synchronize the scaling of clock frequency and supply voltage. The ring oscillator VCO models the critical path of the digital logic in target power domain. Separate regulators are used for digital logic and VCO to isolate the PLL from the digital noise. The PMUs for the RISC and the VS covers from 89MHz to 200MHz with a 1MHz step or from 1.0V to 1.8V with 7.2mV step. The PMU for the RE covers from 22MHz to 50MHz with a 250kHz step with the same voltage characteristics as the other domains. Since all of the blocks are designed with fully static circuits, the blocks can keep running during the frequency and voltage change. Each 0.45mm^2 PMU consumes less than 5.1mW. Figure 15.5.4 shows the DVFS waveforms with a voltage regulation speed of 9mV/1µs.

The chip is fabricated using 0.18µm 6M CMOS technology. The core size is 17.2mm^2, and has 1.57M transistors and 29KB SRAM. Its maximum speed is 141Mvertices/s and power consumption is 52.4mW at 60 fps. Figure 15.5.5 compares the performance of VS with that of other chips, and shows a maximum 3.5× improvement. Figure 15.5.6 summarizes the features of the chip and the chip micrograph is shown in Fig. 15.5.7.

References:

[1] R. Woo, et al., "A 210mW Graphics LSI Implementing Full 3D Pipeline with 264Mtexels/s Texturing for Mobile Multimedia Applications," *ISSCC Dig. Tech. Papers*, pp. 44-45, Feb., 2003.

[2] J. Sohn, et al., "A 50Mvertices/s Graphics Processor with Fixed-Point Programmable Vertex Shader for Moblie Applications," *ISSCC Dig. Tech. Papers*, pp. 192-193, Feb., 2005.

[3] C.-H. Yu, et al., "A 120Mvertices/s Multi-threaded VLIW Vertex Processor for Mobile Multimedia Applications," *ISSCC Dig. Tech. Papers*, pp. 408-409, Feb., 2006.

[4] B.-G. Nam, et al., "A Low-Power Unified Arithmetic Unit for Programmable Handheld 3-D Graphics Systems," *Proc. CICC*, pp. 535-538, Sept., 2006.

[5] T. Fujiyoshi, et al., "An H.264/MPEG-4 Audio/Visual Codec LSI with Module-Wise Dynamic Voltage/Frequency Scaling," *ISSCC Dig. Tech. Papers*, pp. 132-133, Feb., 2005.

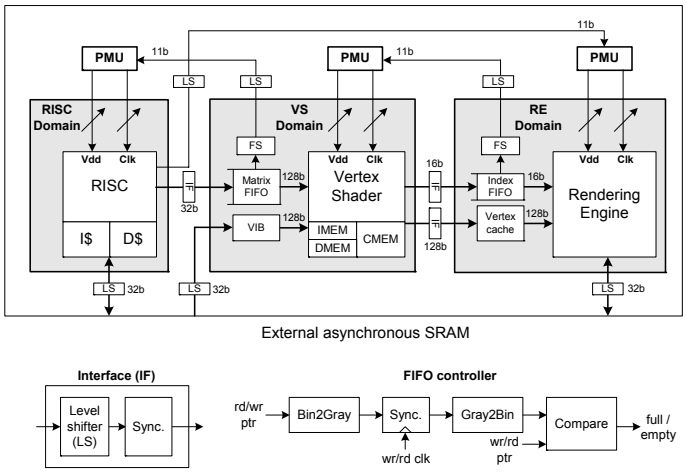

Figure 15.5.1: Overall block diagram of the graphics processor.

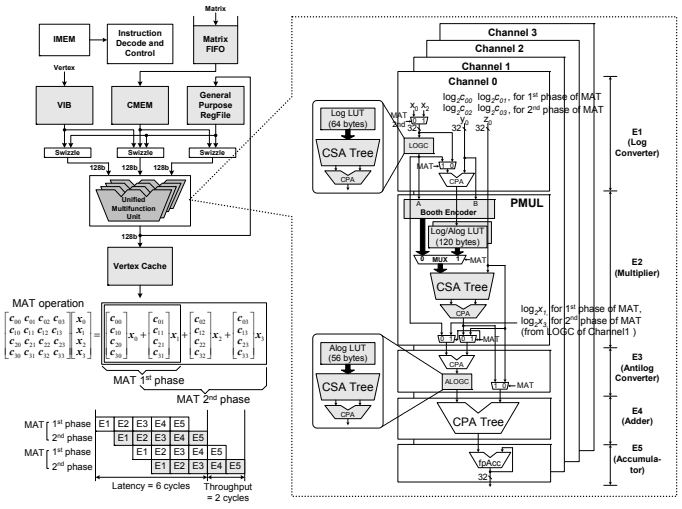

Figure 15.5.2: Vertex shader with unified functional unit.

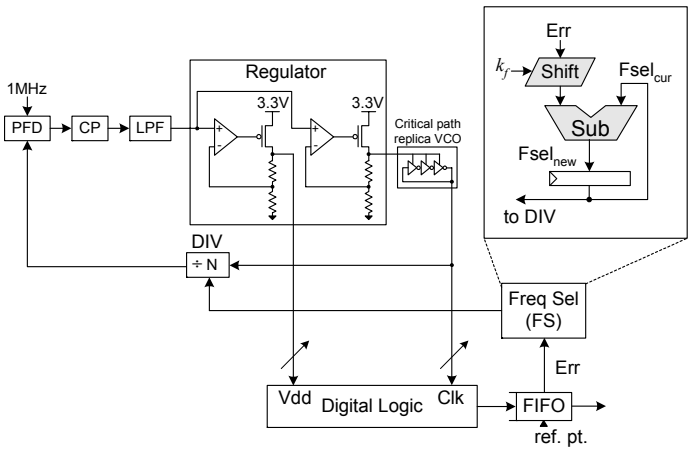

Figure 15.5.3: Power management loop.

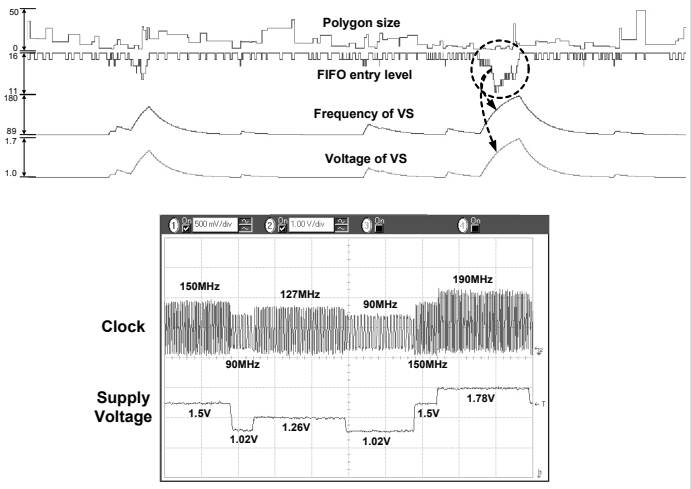

Figure 15.5.4: DVFS operation and waveforms of PMU.

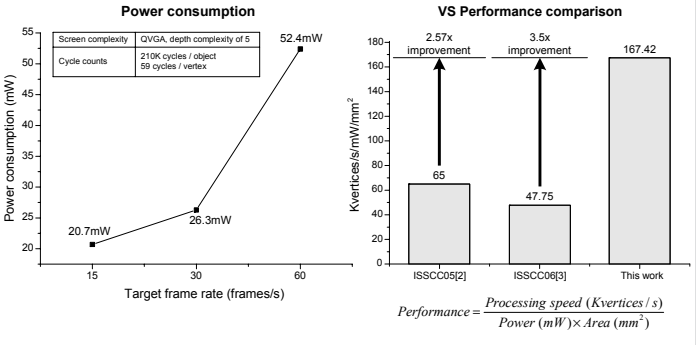

Figure 15.5.5: Power consumption and vertex shader performance comparison.

Process technology	TSMC 0.18um 1-poly 6-metal CMOS technology	
Die size	Core : 17.2mm^2 (9.7mm^2 for VS) Chip : 25mm^2	
Power supply	1.0V - 1.8V for core, 3.3V for I/O	
Operating frequency	RISC : 89MHz - 200MHz VS : 89MHz - 200MHz RE : 22MHz - 50MHz	
Transistor counts	1.6M Logic (968K for VS) 29Kbyte SRAM	
Power consumption	52.4mW at 60fps with QVGA 153mW at full speed (86.8mW for VS)	
Performance	Host	200MIPS
	Geometry	141Mvertices/s (Geometry transformation)
	Rendering	50Mpixels/s, 200Mtexels/s (Bilinear MIPMAP texture filtering)
Features	Standards	OpenGL-ES 2.0 Shader Model 3.0 (w/o texture lookup instruction)
	Power Management	Multiple DVFS power domains - 3 power domains with DVFS - 1 fixed power domain for PMU

Figure 15.5.6: Chip specifications and features.

Continued on Page 603

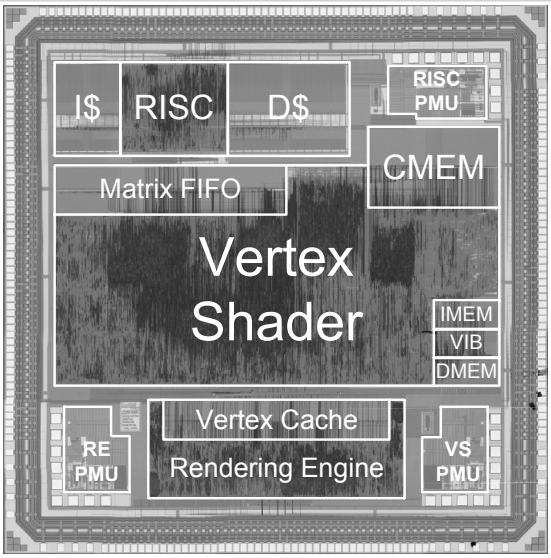

Figure 15.5.7: Chip micrograph.

